Small Neural Networks as Models of Cross-Linguistic Speech Perception

Annika Shankwitz
University of Maryland
ashankwi@umd.edu

Abstract

Given a listener's native language, some non-native contrasts may be harder to discriminate than others. The computation required to mimic this variable difficulty is not yet known. The present work approaches this question by training small supervised feedforward neural networks to perform Spanish vowel classification and then evaluating model classification of Catalan vowels, thereby approximating Spanish-listeners' cross-linguistic perception of Catalan. Vowels were extracted from Spanish and Catalan audio corpora, respectively. Ultimately, Spanish models exhibited expected misperception of Catalan's /e/~/ɛ/, /o/~/ɔ/, and /ɛ/~/a/ contrasts; Spanish-dominant listeners have difficulty perceiving these contrasts, and Spanish models classified Catalan /ɛ/ as /e/ or /a/, and Catalan /ɔ/ as /o/. This demonstrates that small supervised neural models are capable of making specific, cross-linguistic perceptual predictions given realistic input.

1 Introduction

When adults begin to learn a second language (L2), their perception of L2 speech sounds is influenced by their native language (L1); they perceive their L2 "through" their L1 (Caramazza et al., 1973). As such, given a listener's L1, certain L2 speech sounds may be harder to distinguish than others. The computational capacity required to mimic this L1 filter, as well as the type of learning this L1 filter arises from, remain open questions. In other words, when implementing computational models, it is unclear what model size or learning style best captures the behavior of an L1 filter. It is then worthwhile to consider the minimally sufficient combination of these factors that successfully mimics the L1 filter, especially with regard to computational capacity, as doing so enables efficient perceptual predictions to be made and restricts which explanations may be given for L1 filter itself.

Many attempts have been made to model cross-linguistic speech perception, including theoretical (Best, 1995; Flege, 1995; Best and Tyler, 2007) and computational models (see Adriaans, 2024, for an overview). A subset of the latter category are neural-network-based models, which include small recurrent models that learn from laboratory speech (Keidel et al., 2003; Keidel, 2007) as well as larger models commonly used in speech technology that learn from audio corpora (Schatz and Feldman, 2018; Matusevych et al., 2023; Rodriguez et al., 2024). However, it remains unclear if small models, trained and evaluated on audio corpora, can mimic cross-linguistic speech perception.

In this paper, I show that small neural models can (1) learn from realistic input, and (2) make perceptual predictions over entire vowel systems. I focus specifically on modeling the perception of Catalan vowels by Spanish-dominant adult listeners, as this scenario has been described in experimental literature, but has not been fully investigated via modeling. To do so, small four layer feedforward neural networks were trained to perform Spanish vowel classification, and subsequently evaluated on their classification of Catalan vowels. Spanish models misclassified vowels within contrasts that Spanish-dominant listeners have difficulty perceiving, namely /e/~/ɛ/, /o/~/ɔ/, and /ɛ/~/a/. Specifically, Spanish models classified Catalan /ɛ/ as /e/ or /a/, and Catalan /ɔ/ as /o/. This demonstrates that small supervised neural models are capable of mimicking the L1 filter, and show promise as compact models of cross-linguistic speech perception.

2 Neural Models of Cross-Linguistic Speech Perception

Two previous neural modeling efforts are relevant to the present work. Keidel (2007) showed that small models are capable of mimicking the L1 filter, but evaluated model performance on a limited

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 15–24

amount of laboratory speech. Matusevych et al. (2023) suggested that supervised models may correctly predict Spanish listeners' perception of Catalan's /e/~/ɛ/ contrast, but did not carry out a complete analysis. The present work aims to address these limitations.

Keidel (2007)'s model aimed to replicate English speakers' perception of Zulu consonantal contrasts. To impart L1 knowledge of English, the model was trained on English CV syllables[1]. Cross-linguistic perception of Zulu was then determined through evaluation on three Zulu contrasts: /b/~/ɓ/, /kʰ/~/k'/, /ɬ/~/ɮ/.

The implemented model was a five layer recurrent neural network, which was trained to receive cochleagram representations of laboratory-recorded syllables and output vector representations of the input consisting of binary acoustic features. The input layer had 26 units, one for every filter band in the cochleagram representation. The first hidden layer had 50 units, and was recurrently connected to a second 30-unit hidden layer. The first hidden layer was then connected to the 19-unit output layer, where each unit corresponded to a binary acoustic feature (e.g., burst, periodic, F1 low, etc.). Before producing the final output, the output layer used a recurrently connected 20-unit layer to "clean" its response.

The model was evaluated in two ways. Phoneme identification was computed by finding the English segment[2] with the lowest Euclidean distance to the model's output vector. Discrimination was computed by simulating an AXB task using Euclidean distance over the model's hidden state representations for A, X, and B.

Keidel (2007)'s model was implemented with the intention of modeling Best et al. (2001)'s findings on English speakers' perception of Zulu contrasts. In a behavioral study, Best et al. (2001) found that Zulu /b/~/ɓ/ map onto English /b/, /kʰ/~/k'/ map onto English /k/, /ɬ/ maps to voiceless fricatives/affricates involving the tongue tip/body, and /ɮ/ maps to voiced fricatives/affricates involving the tongue tip/body, often including /l/. Additionally, they found that /ɬ/~/ɮ/ is easier to discriminate than /kʰ/~/k'/, which is easier to discriminate than /b/~/ɓ/. Keidel (2007) intended to evaluate model performance on these contrasts, however, his informant produced different variants of these contrasts, namely /ɬ/~/ɮ/, /p/~/ɓ/, and /g/~/k'/.

Ultimately, Keidel (2007)'s model mapped /ɬ/~/ɮ/ onto English /ʃ/~/ʒ/, /g/~/k'/ onto English /g/, and /p/~/ɓ/ onto English /b/. Additionally, /ɬ/~/ɮ/ was the easiest to discriminate, followed by /g/~/k'/ and /p/~/ɓ/, although the difference between the later was not significant. Allowing for differences in stimuli, these findings are largely consistent with Best et al. (2001)'s results.

While Keidel (2007) demonstrates that small models are capable of mimicking the L1 filter, it is unclear how robust this result is. Keidel (2007)'s model was trained and evaluated on a limited amount of laboratory speech; eight speakers produced training syllables, and one speaker produced evaluation syllables. It is then unclear if Keidel (2007)'s approach generalizes to different audio data, or a larger number of speakers. As such, the present work aims to investigate the performance of small models on audio corpora, a different type of audio data[3] containing many speakers.

More recent work considers whether large neural models, particularly those drawn from speech technology applications, are capable of mimicking the L1 filter. One such work, Matusevych et al. (2023), examined whether large autoencoder speech models exhibit the same perceptual difficulties as infants. To do so, they evaluated model discrimination of three contrasts which infants are unable to discriminate, one being Catalan /e/~/ɛ/ for Spanish-learning infants.

Matusevych et al. (2023) assessed the performance of four variations of the autoencoder architecture. Models were evaluated with a machine ABX task (Schatz et al., 2013), where models trained on audio corpora containing target contrasts were expected to achieve lower error rates than models trained on audio corpora not containing target contrasts. For example, Catalan's /e/~/ɛ/ contrast should be more accurately discriminated by a model trained on Catalan than a model trained on Spanish. Three architectures correctly predicted discrimination of at least one contrast, but no model correctly predicted perception of Catalan /e/~/ɛ/.

To examine if model failure on Catalan /e/~/ɛ/ was caused by noising training or evaluation data, Matusevych et al. (2023) trained two supervised phoneme identifiers (Schatz et al., 2021), one on

[1] Consonants included all English stops, fricatives, liquids, along with /m/, and vowels included /a eˡ i oʷ u/.

[2] A specific target vector was created for each English segment.

[3] Keidel (2007)'s speakers produced only syllables. When using audio corpora, target speech sounds are extracted from long, read passages.

their Spanish audio corpus and one on their Catalan audio corpus. They found that the Catalan model achieved significantly lower error rates on /e/~/ɛ/ than the Spanish model, and suggest that a supervised model may successfully model Spanish listeners' perception of Catalan /e/~/ɛ/.

While Matusevych et al. (2023) suggest that Spanish listeners' perception of Catalan /e/~/ɛ/ may be captured by a supervised model, they do not conduct a detailed analysis. As such, the present work aims to carry out this suggestion.

In sum, both Keidel (2007) and Matusevych et al. (2023) offer findings which are limited in some way; Keidel (2007) provides support for small neural models, but said models were trained and evaluated on limited data; Matusevych et al. (2023) provide support for supervised models, but said suggestion is not fully analyzed. Additionally, to the best of my knowledge, neural models have not yet been evaluated on Catalan vowel contrasts other than /e/~/ɛ/. The present study investigates whether small supervised neural networks, trained and evaluated on vowels extracted from audio corpora, successfully mimic Spanish-dominant listeners' perception of the Catalan vowel inventory.

3 A Case Study on Spanish-dominant Spanish-Catalan Bilingual Adults

The present work focuses on Spanish-dominant listeners' perception of Catalan vowels. This pairing was chosen due to its well-defined nature; the vowel inventories of Spanish and Catalan are small enough to be investigated in their entirety, but large enough to offer multiple contrasts of interest. Additionally, this pairing has been explored in experimental work, allowing model classification patterns to be compared to human perceptual patterns.

Spanish has five vowels: /i e a o u/ (Ronquest, 2018), while Catalan has seven vowels[4]: /i e ɛ a o ɔ u/ (Carbonell and Llisterri, 1999). The differing structure of these inventories leads to perceptual difficulties for Spanish-dominant listeners.

Adult Spanish-dominant bilinguals perceive certain Catalan vowel contrasts differently than Catalan-dominant bilinguals. Relative to their Catalan-dominant counterparts, Spanish-dominant bilinguals take longer to identify, and achieve lower discrimination accuracy on Catalan's /e/~/ɛ/ and /o/~/ɔ/ contrasts (Pallier et al., 1997; Sebastián-Gallés and Soto-Faraco, 1999; Amengual, 2016). Spanish-dominant bilinguals have also been shown to achieve lower discrimination accuracy on Catalan's /ɛ/~/a/ contrast (Amengual, 2016). Perceptual differences, however, do not exist for all Catalan vowel contrasts; Spanish-dominant and Catalan-dominant bilingual adults achieve comparable discrimination accuracy on Catalan's /i/~/e/, /u/~/o/, and /ɔ/~/a/ contrasts (Amengual, 2016).

These perceptual patterns can be used to outline expected model behavior. The contrasts known to be difficult for Spanish-dominant adult bilinguals, /e/~/ɛ/, /o/~/ɔ/, and /ɛ/~/a/, all involve one segment which does not appear in Spanish: /ɛ/ or /ɔ/. Recall that Spanish models are trained to perform Spanish vowel classification, and as a result, are only capable of producing the labels /i e a o u/. As such, /ɛ/ should be classified as Spanish /e/ or /a/ and /ɔ/ should be classified as Spanish /o/.

4 Methods

The project's code[5] and forced alignments[6] are publicly available.

4.1 Data

Spanish. Spanish vowels were extracted from the Spanish partition of the Multilingual Librispeech Corpus, a large audio corpus of read speech (Pratap et al., 2020). The Spanish partition contains roughly 1000 hours of speech, split across training (918 hrs), development (10 hrs), and testing (10 hrs) sets.

Catalan. Catalan vowels were extracted from the Catalan partition of the Open-Source High Quality Speech Datasets for Basque, Catalan, and Galician, a small audio corpus of read speech (Kjartansson et al., 2020). The Catalan partition contains roughly 9.25 hours of speech.

4.2 Data Preprocessing

Audio data was preprocessed to reflect human auditory processing for vowels over short time intervals.

Forced Alignment. Audio corpora were aligned using the Montreal Forced Aligner (McAuliffe et al., 2017). Spanish audio files were aligned

[4] Not including [ə].

[5] https://github.com/its-Annika/neural-speech-perception. The coding portion of this project was completed with assistance from OpenAI's ChatGPT. The scope of this assistance is described in the README file of the project's repository.

[6] https://github.com/its-Annika/es_ca_alignments

with provided Spanish acoustic and grapheme-to-phoneme models (McAuliffe and Sonderegger, 2022a,b); Catalan audio files were aligned with Vargo (2025)'s custom Catalan bundle[7, 8].

Sampling. After corpora were aligned, random sampling was used to produce model training, development, and evaluation datasets. For Spanish, 50 random productions of /i e a o u/ were sampled per speaker. Spanish training, development, and testing sets were formed by sampling from the corresponding splits in the Spanish partition.

For Catalan, 30 random productions of /i e ɛ a o ɔ u/ were sampled per speaker. Given the Catalan corpus's small size, vowels were split into only training and testing sets, and cross-validation was used in model training. The Catalan training set was formed by sampling from 80% of speakers in the Catalan partition. The Catalan testing set was formed by sampling from the remaining 20% of speakers. Catalan speakers were divided randomly.

Speakers who did not produce enough instances of each vowel were not included in final data files. Additionally, vowels less than 30 ms in duration were not eligible for sampling (see Slicing). A summary of these files can be seen in Table 1.

File	Total Vowels	Samples by Vowel	Speakers
Spanish training	18,750	3750	75 (M=34)
Spanish development	5000	1000	20 (M=10)
Spanish testing	5000	1000	20 (M=10)
Catalan training	5880	840	28 (M=13)
Catalan testing	1680	240	8 (M=3)

Table 1: Data file details.

Slicing. Sampled vowels' start and end points were extracted from textGrids produced during alignment using `praatIO` (Mahrt, 2016). A 30 ms slice was then taken from sampled vowels, centered on the midpoint. At a 16 kHz sampling rate, vowel slices contained 480 time steps. Before slicing, audio files containing a sampled vowel were normalized to -25 dB LUFS using `pyloudnorm` (Steinmetz and Reiss, 2021).

Conversion to Cochleagrams. Following Keidel (2007), vowel slices were converted to cochleagram representations, representations which approximate how sound is received and processed by the human cochlea. Conversion was done using `pycochleagram` (Gonzalez, 2018). Representations used 24 filter bands to capture frequencies from 50 to 8000 Hz (roughly one band per 1/3 octave), where frequencies captured by individual bands overlapped by 50%, along with a low-pass and high-pass filter. Conversion resulted in representations of shape 480 time steps x 26 filter bands. The time dimension was then removed by averaging over said dimension, resulting in 1x26 vectors.

4.3 Model Architecture and Training

While Keidel (2007) implemented a recurrent neural network, the current work implements feedforward neural networks. This change was made for two reasons: a time dimension is not necessary to represent monophthongal vowels, rendering recurrence unnecessary, and feedforward networks are simpler than recurrent networks, allowing the performance of models simpler than Keidel (2007)'s to be examined. Additionally the present models differ in training objective from Keidel (2007)'s model; Keidel (2007)'s model learned to produce acoustic feature vectors, while the present models learn to produce a single vowel label. This change was made so as to not restrict models' learned representations to a specific, provided feature set. Two types of models were trained: a Catalan baseline model, intended to simulate Catalan-dominant listeners (which produces the labels /i ɛ e a o ɔ u/) and Spanish models, intended to simulate Spanish-dominant listeners (which produce the labels /i e a o u/).

Models are four layer feedforward neural networks with two hidden layers[9]. ReLU non-linear activation is applied after both hidden layers. The input layer has 26 units (one unit for each filter band) and the output layer has 5 or 7 units (one

[7] A small portion of both corpora were unable to be aligned. Spanish: 2,194 files, Catalan: 102 files.

[8] Before alignment, probabilistic information was removed from Vargo (2025)'s dictionary as it contained both probabilistic and non-probabilistic entries.

[9] The performance of five layer networks with three hidden layers was also explored. The classification accuracy of five layer models on their L1 (Catalan for the baseline, Spanish for Spanish models), either did not improve or degraded relative to the accuracies of four layer models. As such, this work focuses on four layer models.

unit for each Spanish or Catalan vowel). This general architecture can be seen in Figure 1. Models were trained using cross entropy loss and the Adam optimizer (Kingma and Ba, 2015).

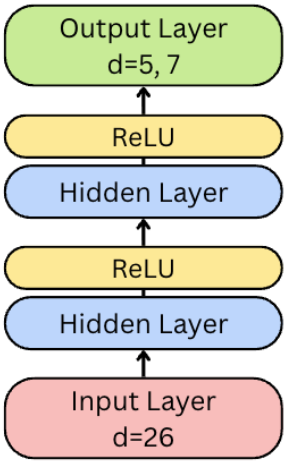

Figure 1: Model architecture.

Catalan Baseline. To provide a comparison for Spanish models, a Catalan baseline model was implemented. Given the small size of the Catalan training set, the model was implemented with a combination of grid search and 5-fold cross-validation. Grid search was conducted targeting the learning rate (0.01, 0.001, 0.0001), hidden layer dimension (7, 14, 26, 52)[10], number of training epochs (10, 20, 30, 40, 50, 60, 70), and batch size (4, 16, 32). 5-fold cross-validation was used to find optimal model parameters, which were then used to train a fresh initialization. Optimal parameters consisted of a 0.001 learning rate, 70 training epochs, a hidden dimension of 52, and a batch size of 4. These parameters achieved an average 66% accuracy during cross-validation.

Spanish Models. To find optimal parameters, grid search was conducted targeting the learning rate (0.01, 0.001, 0.0001), hidden layer dimension (5, 10, 26, 52), number of training epochs (10, 20, 30, 40, 50, 60, 70), and batch size (4, 16, 32). Models were trained on the Spanish training set, and their classification accuracy was evaluated on the Spanish development set.

Two models were chosen for evaluation. The first was the best model found during grid search. This model had parameters of 0.001 learning rate, 50 training epochs, a hidden dimension of 52, and a batch size of 4, and achieved 76.86% on the Spanish development set (henceforth Spanish-52). The second model was the highest performing model with a hidden dimension of 5 units (henceforth Spanish-5). This model had additional parameters of 0.01 learning rate, 30 training epochs, and a batch size of 16, and achieved 73.22% accuracy on the Spanish development set. An overview of evaluated models can be seen in Table 2.

	Spanish-5	Spanish-52	Baseline
Learning Rate	0.01	0.001	0.001
Training Epochs	30	50	70
Hidden Dimension	5	52	52
Batch Size	16	4	4
Accuracy	73.22%	76.86%	66%

Table 2: Overview of evaluated models.

4.4 Evaluation Metrics

Models were evaluated in two ways. First, model categorization of vowels specific to Catalan, namely /ɛ/ and /ɔ/, was measured. Given empirical results, Spanish models were expected to classify /ɛ/ as /e/ or /a/, and /ɔ/ as /o/, more frequently than the baseline (Pallier et al., 1997; Sebastián-Gallés and Soto-Faraco, 1999; Amengual, 2016).

Second, model discrimination was measured for three test contrasts, /e/~/ɛ/, /o/~/ɔ/, /ɛ/~/a/, along with three control contrasts /i/~/e/, /u/~/o/, /ɔ/~/a/. Discrimination was approximated by comparing a model's label distributions for two vowels within a contrast. For example, consider the /e/~/ɛ/ contrast. A Catalan model is expected to classify /ɛ/ and /e/ as two separate vowels, and as such, the labels assigned to /ɛ/ and /e/ should not have similar distributions. A Spanish model, however, is expected to classify /ɛ/ and /e/ as the same vowel, and as such, the labels assigned to /ɛ/ and /e/ should have similar distributions. Total variation distance (1) was used for this purpose, where P and Q represent probability distributions over a model's label set[11].

[10] 7 units corresponds to one unit for each vowel in the Catalan inventory; 26 units corresponds to one unit for each filter band in cochleagram representations.

[11] The baseline produces distributions over seven categories while Spanish models produce distributions over only five categories. This introduces a potential confound when comparing

$$\mathrm{TV}(\mathrm{P}, \mathrm{Q}) = \frac{1}{2} \sum_{x} |\mathrm{P}(x) - \mathrm{Q}(x)| \qquad (1)$$

Differences between model total variation distances were compared to Amengual (2016)'s findings. Catalan and Spanish models were expected to exhibit similar total variation distances on control contrasts, while Spanish models were expected to show lower total variation distances than the Catalan baseline on test contrasts.

5 Results

To determine whether models learned their L1, the baseline was evaluated on the Catalan testing set, and Spanish models were evaluated on the Spanish testing set. Heatmaps for the baseline, Spanish-5, and Spanish-52 appear in Figures 2, 3, and 4, respectively.

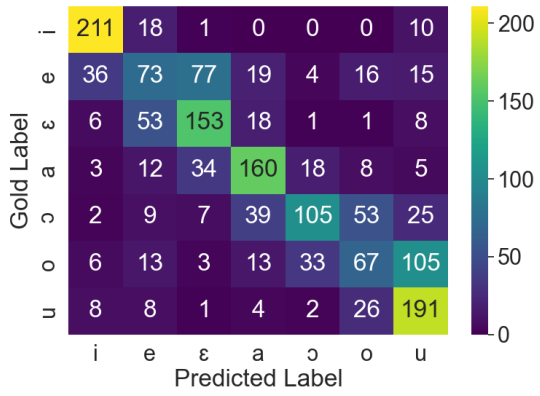

Figure 2: Baseline performance on Catalan testing set. Accuracy: 57.14%

Given the Catalan training set's small size, it is unsurprising that the baseline did not completely learn the Catalan inventory, that is, that a perfect diagonal does not appear in Figure 2. Despite this, the model still provides a comparison point for Spanish models; the baseline has knowledge of the two vowels specific to the Catalan inventory, /ɛ/ and /ɔ/, and the distribution of the baseline's predictions can be compared to those of the Spanish models.

Both Spanish models successfully learned the Spanish vowel inventory, as evidenced by the bright diagonals in Figures 3 and 4. These models can then be used to approximate Spanish-dominant listeners' perception of Catalan vowels.

Recall that the three Catalan contrasts of interest are /e/~/ɛ/, /o/~/ɔ/, and /ɛ/~/a/. Spanish models are expected to classify /ɛ/ as /e/ or /a/, and /ɔ/ as /o/. To evaluate Spanish models' classification of Catalan vowels, models were evaluated on the Catalan testing set. Spanish-5's classification heatmap can be seen in Figure 5 and Spanish-52's classification heatmap in Figure 6.

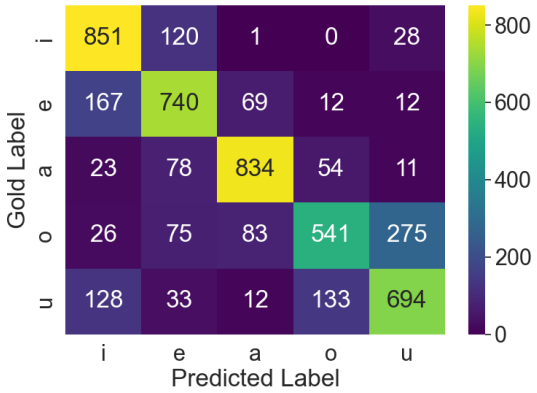

Figure 3: Performance of Spanish-5 on Spanish testing set. Accuracy: 73.2%

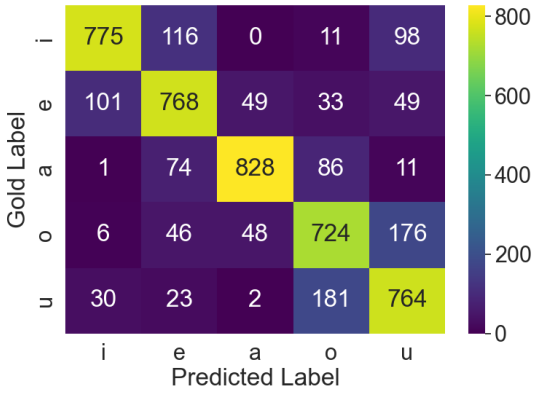

Figure 4: Performance of Spanish-52 on Spanish testing set. Accuracy: 77.18%

Figures 5 and 6 indicate that Spanish models largely maintain knowledge of vowels shared between the Spanish and Catalan inventories. Additionally, in line with predictions, /ɛ/ is most frequently classified as /e/, followed by /a/, and /ɔ/ is most frequently classified as /o/.

A more concrete evaluation of model classification can be achieved by investigating model classification percentages. Percentages within Table 3

between models, as total variation distances for the Catalan baseline may be artificially larger than Spanish models' differences. In practice, however, Catalan and Spanish models have similar total variation distance on control contrasts when hidden dimension is held constant.

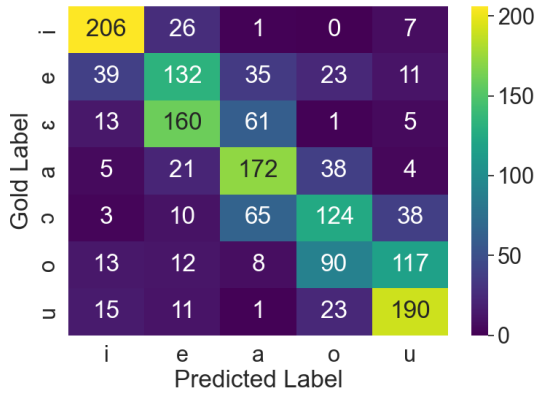

Figure 5: Performance of Spanish-5 on Catalan testing set. Accuracy: 47.02%

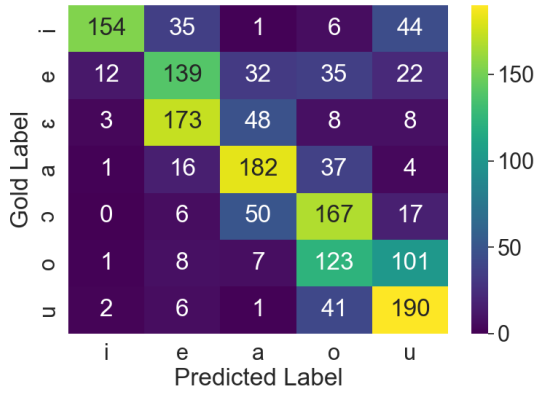

Figure 6: Performance of Spanish-52 on Catalan testing set. Accuracy: 46.9%

indicate that Spanish models did indeed classify /ɛ/ as /e/ or /a/, and /ɔ/ as /o/, more frequently than the baseline, with the difference in classification percentage being most pronounced for /ɛ/ as /e/, followed by /ɔ/ as /o/, and finally /ɛ/ as /a/. This demonstrates that Spanish models classified /ɛ/ and /ɔ/ differently than the Catalan baseline, and in ways that align with empirical results.

	/ɛ/ as /e/	/ɛ/ as /a/	/ɔ/ as /o/
Spanish-5	66.67%	25.42%	51.67%
Spanish-52	72.08%	20.0%	69.58%
Baseline	22.08%	7.5%	22.08%

Table 3: Frequency with which models classify /ɛ/ as /e/, /ɛ/ as /a/, and /ɔ/ as /o/.

Direct comparison to experimental results is available for one contrast: /e/~/ɛ/. Pallier et al. (1997) assessed Catalan-dominant and Spanish-dominant bilinguals' identification of a synthetic vowel continuum spanning from [e] to [ɛ], and found that Catalan-dominant identifications formed a clear s-shaped curve, while Spanish-dominant identifications remained largely flat across the vowel continuum. Participant identifications of the first and last stimuli (prototypical [e] and [ɛ], respectively) as /ɛ/ can be seen in Table 4.

	[e] as /ɛ/	[ɛ] as /ɛ/
Catalan-dominant	~15%	~82%
Spanish-dominant	~55%	~35%

Table 4: Pallier et al. (1997)'s classification percentages for stimulus 1 - [e], and stimulus 7 - [ɛ]. Percentages are estimated from Pallier et al. (1997) Figure 1.

Using the percentages in Table 4, other identifications can be inferred; Catalan-dominant participants identified ~18% of [ɛ]s as /e/; Spanish-dominant participants identified ~65% of [ɛ]s as /e/. The current models mirror this pattern, with the Catalan baseline classifying 22.08% of [ɛ]s as /e/, and the Spanish models classifying upwards of 66.67% of [ɛ]s as /e/. This demonstrates that models' identification of Catalan /ɛ/ aligns with that of the population they intend to capture.

Models were also evaluated on their discrimination, which was approximated using total variation distance, and compared to empirical results. In an AXB task, Amengual (2016) found that Spanish-dominant listeners show comparable discrimination accuracy to Catalan-dominant listeners on Catalan's /i/~/e/, /u/~/o/, /ɔ/~/a/ contrasts, but significantly higher discrimination error than Catalan-dominant listeners on Catalan's /e/~/ɛ/ (Catalan-dominant 5%, Spanish-dominant 11%), /o/~/ɔ/ (Catalan-dominant 4%, Spanish-dominant 12.3%), /ɛ/~/a/ (Catalan-dominant 6.2%, Spanish-dominant 11.2%) contrasts. As such, Spanish models are expected to exhibit similar total variation distances to the baseline on /i/~/e/, /u/~/o/, /ɔ/~/a/, but lower total variation distances than the baseline on /e/~/ɛ/, /o/~/ɔ/, /ɛ/~/a/. Model total variation distances can be seen in Table 5.

Comparing the Catalan baseline and Spanish-52 partially yields the same results found by Amengual (2016); the models show similar total variation distances on the control contrasts, and Spanish-52 shows lower total variation distance than the baseline on one test contrast, namely /e/~/ɛ/.

Comparing the Catalan baseline and Spanish-5 does not yield the same results found by Amengual (2016); Spanish-5 shows lower total variation

	Test Contrasts			Control Contrasts		
	/e/~/ɛ/	/ɛ/~/a/	/o/~/ɔ/	/i/~/e/	/u/~/o/	/ɔ/~/a/
Spanish-5	0.22	0.62	0.38	0.7	0.31	0.5
Spanish-52	0.21	0.68	0.36	0.68	0.37	0.6
Baseline	0.32	0.69	0.42	0.73	0.37	0.63

Table 5: Total variation distance between model classification distributions.

distances than the baseline on test conditions as well as control conditions. So, while Spanish-5's classification of Catalan vowels is consistent with Spanish-dominant listeners' perception, this is not reflected in the model's discrimination.

In sum, Spanish models' classification differed from that of the Catalan baseline; Spanish-5 and Spanish-52 classified /ɛ/ as /e/ or /a/, and /ɔ/ as /o/ more frequently than the baseline. Additionally, while Spanish models largely did not exhibit expected discrimination results, Spanish-52 did show lower total variation distance on Catalan's /e/~/ɛ/ contrast than the baseline. Ultimately, these perceptual patterns exhibited by Spanish models are consistent with those demonstrated by adult Spanish-dominant listeners.

6 Discussion

In an effort to model Spanish-dominant listeners' cross-linguistic perception of Catalan vowels, this work trained small supervised feedforward neural networks to perform Spanish vowel classification and evaluated model classification of Catalan vowels. Vowels were extracted from Spanish and Catalan audio corpora, and processed to represent human auditory processing of vowels. While Spanish models largely did not exhibit expected discrimination patterns, Spanish models classified Catalan /ɛ/ as /e/ or /a/, and Catalan /ɔ/ as /o/, closely mimicking perceptual patterns exhibited by adult Spanish-dominant listeners.

The present work builds on Keidel (2007)'s recurrent neural model of cross-linguistic speech perception by mimicking L2 perceptual patterns using simpler networks trained and evaluated on larger amounts of data. Keidel (2007)'s model was a five layer recurrent neural network with hidden layers as large as 50 units, while the largest Spanish model had a comparable hidden dimension but a simpler, four layer feedforward architecture. Moreover, even a network with 5-unit hidden layers was able to mimic Spanish-dominant listeners' perception. Additionally, Keidel (2007)'s model was trained and evaluated on a limited amount of laboratory speech, while present models were trained and evaluated on vowels extracted from audio corpora. In short, this work demonstrates that models simpler than Keidel (2007)'s can predict L2 perceptual patterns, even given more complicated data.

This work also builds upon Matusevych et al. (2023) by showing that supervised models are capable of capturing Spanish-dominant listeners' perception of Catalan's /e/~/ɛ/ contrast. This suggests that learning style, instead of model size, is more important when modeling cross-linguistic speech perception, as Matusevych et al. (2023)'s large unsupervised models failed to predict Spanish-dominant listeners' perception of this contrast.

Additionally, this work demonstrates that computational models of cross-linguistic speech perception need not necessarily borrow from speech technologies. Recent modeling work has focused on using real speech, namely audio corpora, to train computational models of cross-linguistic speech perception. As such, model size has increased, with large speech models often being employed (Schatz and Feldman, 2018; Matusevych et al., 2023). The present work, however, illustrates that realistic input does not necessarily require large, specialized models; small feedforward networks are able to learn from audio corpora, and make predictions of cross-linguistic speech perception. This is advantageous, as small models are generally lighter to implement and easier to interpret than larger models. Small models also often make use of simpler mechanisms, which may be desirable from a cognitive perspective.

This work is, of course, not without limitations. One straightforward limitation lies in the baseline, specifically in the small amount of data used to train it. The baseline was likely under-trained, making it an imperfect point of comparison.

This work also implemented only a single model architecture, the feedforward neural network. Given this work's interest in small, simple models, the performance of approaches such as k-nearest

neighbors, k-means clustering, or logistic regression, should also be investigated.

A question raised by this work concerns model perception of /u/~/o/. All models classified /o/ as /u/ with some frequency[12]. This is interesting, given that balanced Spanish-Catalan bilingual infants have been shown to briefly lose the ability to discriminate the /u/~/o/ contrast (Bosch and Sebastián-Gallés, 2003). While the present work did not focus on this contrast, future work should more carefully consider model perception of /u/~/o/.

Finally, to explore the generalizability of this approach, the present work should be replicated on different segmental contrasts or language pairs. For example, an extension within the Spanish-Catalan pairing is the /ʃ/~/ʒ/ contrast, which is present in Catalan, but not in Spanish (Sebastián-Gallés and Soto-Faraco, 1999).

7 Acknowledgments

I am grateful to Bill Idsardi and Naomi Feldman for their valuable feedback during the course of this project. I also thank the anonymous reviewers for their helpful comments.

References

Frans Adriaans. 2024. Computational approaches to bilingual phonetics and phonology. In Mark Amengual, editor, *Bilingual Phonetics and Phonology*, pages 126–144. Cambridge University Press.

Mark Amengual. 2016. The perception and production of language-specific mid-vowel contrasts: Shifting the focus to the bilingual individual in early language input conditions. *International Journal of Bilingualism*, 20(2):133–152.

Catherine Best. 1995. A direct realist view of cross-language speech perception. In Winifred Strange, editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 171–204. Timonium, MD: York Press.

Catherine Best, Gerald McRoberts, and Elizabeth Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2):775–794.

Catherine Best and Michael Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In Ocke-Schwen Bohn and Murray Munro, editors, *Language experience in second language speech learning: In honor of James Emil Flege*, pages 13–34. John Benjamins.

Laura Bosch and Núria Sebastián-Gallés. 2003. Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, 46:217–243.

Alfonso Caramazza, Grace Yeni-Komshian, E Zurif, and E Carbone. 1973. The acquisition of a new phonological contrast: the case of stop consonants in french english bilinguals. *The Journal of the Acoustical Society of America*, 54(2):421–428.

Joan Carbonell and Joaquim Llisterri. 1999. Catalan. In *Handbook of the International Phonetic Association*, pages 61–65. Cambridge University Press.

James Flege. 1995. Second language speech learning: Theory, findings and problems. In Winifred Strange, editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 233–277. Timonium, MD: York Press.

Ray Gonzalez. 2018. pycochleagram.

James Keidel. 2007. *Behavioral, Modeling and Neuroimaging Studies of Auditory Expertise*. Ph.D. thesis, University of Wisconsin-Madison.

James Keidel, Jason Zevin, Keith Kluender, and Mark Seidenberg. 2003. Modeling the role of native language knowledge in perceiving nonnative speech contrasts. In *15th International Congress of Phonetic Sciences*, pages 2221–2224, Barcelona, Spain.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 2020. Open-source high quality speech datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages*, pages 21–27, Marseille, France. European Language Resources association.

Tim Mahrt. 2016. PraatIO.

Yevgen Matusevych, Thomas Schatz, Herman Kamper, Naomi Feldman, and Sharon Goldwater. 2023. Infant phonetic learning as perceptual space learning: A crosslinguistic evaluation of computational models. *Cognitive Science*, 47(4):e13314.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017*, pages 498–502, Stockholm, Sweden.

[12]Baseline: 43.75%, Spanish-5 evaluated on Spanish: 27.5%, Spanish-52 evaluated on Spanish: 17.6%, Spanish-5 evaluated on Catalan: 48.75%, Spanish-52 evaluated on Catalan: 42.08%.

Michael McAuliffe and Morgan Sonderegger. 2022a. Spanish MFA acoustic model v2.0.0. Technical report.

Michael McAuliffe and Morgan Sonderegger. 2022b. Spanish (Spain) MFA G2P model v2.0.0. Technical report.

Christophe Pallier, Laura Bosch, and Núria Sebastián-Gallés. 1997. A limit on behavioral plasticity in speech perception. *Cognition*, 64(3):B9–B17.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761. ISCA.

Joselyn Rodriguez, Kamala Sreepada, Ruolan Leslie Famularo, Sharon Goldwater, and Naomi Feldman. 2024. Self-supervised speech representations display some human-like cross-linguistic perceptual abilities. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 458–463, Miami, FL, USA. Association for Computational Linguistics.

Rebecca Ronquest. 2018. Vowels. In Kimberly Geeslin, editor, *The Cambridge Handbook of Spanish Linguistics*, pages 145–164. Cambridge University Press.

Thomas Schatz and Naomi Feldman. 2018. Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception. In *Conference on Cognitive Computational Neuroscience*, Philadelphia, Pennsylvania.

Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7):e2001844118.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In *Interspeech 2013*, pages 1781–1785, Lyon, France.

Núria Sebastián-Gallés and Salvador Soto-Faraco. 1999. Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, 72(2):111–123.

Christian J. Steinmetz and Joshua D. Reiss. 2021. pyloudnorm: A simple yet flexible loudness meter in python. In *150th Audio Engineering Society Convention*.

Julian Vargo. 2025. Catalan montreal forced alignment bundle.